



CREI
Centro di Ricerca
Interdipartimentale
di Economia delle Istituzioni

CREI Working Paper no. 2/2012

**On Baseline conditions for zero-inflated longitudinal
count data**

by

Antonello Maruotti
University of Roma3

Valentina Raponi
University of Rome “La Sapienza”

available online at <http://host.uniroma3.it/centri/crei/pubblicazioni.html>

ISSN 1971-6907

Outputs from CREI research in progress, as well contributions from external scholars and draft reports based on CREI seminars and workshops, are published under this series. Unless otherwise indicated, the views expressed are attributable only to the author(s), not to CREI nor to any institutions of affiliation.

On baseline conditions for zero-inflated longitudinal count data

Antonello Maruotti ^{*}and Valentina Raponi[†]

Abstract

We describe a mixed-effects hurdle model for zero-inflated longitudinal count data, where a baseline variable is included in the model specification. Association between the count data process and the endogenous baseline variable is modeled through a latent structure, assumed to be dependent across equations. We show how model parameters can be estimated in a finite mixture context, allowing for overdispersion, multivariate association and endogeneity of the baseline variable. The model behavior is investigated through a large scale simulation experiment. An empirical example on health care utilization data is provided.

Keywords: Hurdle model - Baseline conditions - Longitudinal count data - Zero-inflation.

1 Introduction

Great attention has recently been devoted to model longitudinal count data in a regression framework. Salient features of longitudinal count data may include, in addition to non-negativity of outcomes, a large fraction of zeros, the presence of association between

^{*}Dipartimento di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre, Roma, Italy, email: antonello.maruotti@uniroma3.it

[†]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Roma, Italy

repeated measures recorded on the same unit and unobserved heterogeneity. Even if the Poisson regression model represents a useful tool for dealing with count data processes, different approaches need to be investigated to account for all data features. Variance components models (see e.g. Aitkin, 1996) can be used to account for unobserved heterogeneity; while dynamics models can be used to capture the dependence across repeated measures (see e.g. Aitkin and Alfò, 2003). Several models have been recently proposed to deal appropriately with the zero-modification (zero-truncated and zero-inflated/deflated) issue, i.e. more zeros than expected under standard count distribution (Min and Agresti, 2005; Hasan and Sneddon, 2009; Greene, 2009; Todem et al., 2010).

In the following we focus on two-part models, also known as hurdle models (Mullahy, 1986), to introduce flexibility by allowing the zeros and the positive outcomes to be generated by two different processes: a binary model generating the zeros and a truncated at zero model generating the conditional distribution of the positives. In such a framework, generally, the two processes are assumed independent or, equivalently, the binary process (describing the *hurdle*) is considered exogenous. The independence assumption allows to factorize the log-likelihood as the sum of two log-likelihood functions (one for each process) and, therefore, the computational burden is dramatically simplified since parameter estimates can be obtained maximizing the two terms separately. This assumption may be too restrictive when dealing with real data (see e.g. Deb et al., 2006; Auteri and Maruotti, 2011) and a model structure allowing for dependence between the processes can be required. If the exogeneity assumption is invalid, parameter estimates from *standard* hurdle models could be inconsistent. Thus, we define a mixed effects hurdle model (Alfò and Maruotti, 2010), relying on a simple association structure, described through the inclusion of random terms in the model specification, which allows for the endogeneity of the *hurdle*. In order to include some sort of serial dependence, we integrate the mixed effects hurdle model by considering the effect of baseline outcomes on subsequent responses, with special focus on short longitudinal datasets. Nevertheless, the baseline itself is subject to the influence of same covariates as the response variable; it is thus not exogenous and is not independent of the random effects in the model. This is

usually known in the literature as endogeneity bias, i.e. the statistical bias that arises when an endogenous variable is treated as exogenous (see e.g. Davidson and Mackinnon, 1993). The issue of endogeneity has long been recognized in econometrics literature, where several techniques for dealing with it exist (see e.g. Alfò et al., 2011). Even if the inconsistency of the maximum likelihood estimation which results from ignoring the baseline disappears as the length of the sequence tends to infinity (Hsiao, 2003), simply ignoring the endogeneity of the baseline may lead to model misspecification and misleading interpretation for the baseline effect.

The importance of modeling the baseline in longitudinal data has been deeply analyzed by Fotouhi (2007; 2005) via simulations, and has a long history in econometrics since the seminal paper of Heckman (1981). In the following two approaches are considered: the naive approach of ignoring the endogenous nature of the baseline variable and the joint approach where the baseline is modeled as an endogenous variable in a mixed-effects hurdle model framework. The impact of baseline outcome on subsequent observations has been evaluated via a large-scale simulation study. The Expectation-Maximization (EM) algorithm for non-parametric maximum likelihood is adopted (see e.g. Alfó and Aitkin, 2006 for further details in a count data framework). We leave the random effects distribution unspecified and show how parameter estimation can be performed in a finite mixture context.

The plan of the paper is as follows. We provide model details in Section 2: the naive and the joint approaches are described as well as the mixed-effects hurdle model and its computational details. A simulation study is conducted in Section 3 in order to understand the model behavior under different sample sizes, panel lengths and random effects distributions. An example on health service utilization is provided in Section 4. A discussion on further developments and drawbacks of the proposed approach concludes.

2 Modeling approach

In this section we discuss the general mixed effects hurdle model specification. The naive and the joint approaches are presented and, thus, a model taking into account the data features and the endogeneity of the baseline is developed. Computational details are provided to obtain maximum likelihood estimates by using an EM-based algorithm.

2.1 Model specification

Consider a count data process Y_{it} recorded on each of n units ($i = 1, \dots, n$) over T time periods ($t = 1, \dots, T$) and two sets of covariates \mathbf{x}_{it1} and \mathbf{x}_{it2} that may or may not be the same. The mixed effects hurdle model postulates a hierarchical structure: first a decision process leading to the choice $y_{it} = 0$ versus $y_{it} > 0$, and then a process accounting for $y_{it} > 0$. The probability function of the observed variable is then given by

$$f(y_{it} | \cdot) = \Pr(Y_{it} = y_{it}) = \begin{cases} \pi_{it}, & y_{it} = 0 \\ (1 - \pi_{it}) \frac{\exp(-\lambda_{it}) \lambda_{it}^{y_{it}}}{y_{it}! (1 - \exp(-\lambda_{it}))}, & y_{it} > 0 \end{cases}$$

where $\pi_{it} = \Pr(y_{it} = 0)$ and λ_{it} represent the canonical parameters for the binary and the count processes, respectively. A generalized linear mixed model is then used to focus on π_{it} and λ_{it} : a logistic model for π_{it} and a log-linear model for λ_{it}

$$\text{logit}(\pi_{it}) = \mathbf{x}'_{it1} \boldsymbol{\beta} + b_{i1} \tag{1}$$

$$\log(\lambda_{it}) = \mathbf{x}'_{it2} \boldsymbol{\gamma} + b_{i2} \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression parameters associated with \mathbf{x}_{it1} and \mathbf{x}_{it2} , respectively; while $\mathbf{b}_i = (b_{i1}, b_{i2})$ denotes a set of unit- and process-specific random effects drawn from a bivariate distribution, $h(\cdot)$.

We aim at extending this approach to include dependence on baseline counts, allowing for a quite general association structure. A common and straightforward strategy of accounting for the baseline, say y_{i0} , in the model specification is to include it in the linear predictors.

In this case, equations (1) and (2) can be rewritten as

$$\text{logit}(\pi_{it}) = \mathbf{x}'_{it1}\boldsymbol{\beta} + \alpha_1 y_{i0} + y_{i0}\mathbf{x}'_{it1}\boldsymbol{\phi}_1 + b_{i1} \quad (3)$$

$$\log(\lambda_{it}) = \mathbf{x}'_{it2}\boldsymbol{\gamma} + \alpha_2 y_{i0} + y_{i0}\mathbf{x}'_{it2}\boldsymbol{\phi}_2 + b_{i2} \quad (4)$$

where $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are vectors of regression parameters associated with interactions between baseline counts and other covariates. Not all these interactions will be taken into account; therefore, some elements of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ can be set equal to zero.

Given the model assumptions, we can write down the conditional distribution of the observed responses for the i -th unit as

$$L_i(y_{i1}, y_{i2}, \dots, y_{iT} | y_{i0}) = \int \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it1}, \mathbf{x}_{it2}, y_{i0}, b_{i1}, b_{i2}) h(\mathbf{b}_i | y_{i0}) d\mathbf{b}_i \quad (5)$$

2.2 The naive approach

The naive approach simply assumes independence between the baseline and the random terms. Therefore, we can integrate function (5) over the distribution of \mathbf{b}_i without the need of specifying the conditional distribution $h(\mathbf{b}_i | y_{i0})$. The resulting integrated conditional likelihood for the i -th unit is given by

$$L_i(y_{i1}, y_{i2}, \dots, y_{iT} | y_{i0}) = \int \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it1}, \mathbf{x}_{it2}, y_{i0}, b_{i1}, b_{i2}) h(\mathbf{b}_i) d\mathbf{b}_i \quad (6)$$

The independence assumption, i.e. assuming the baseline exogenous while it is in fact endogenous, can be too restrictive, even if it reduces the computational burden. A confounding effect may arise: we cannot distinguish between the influence of the baseline on the following counts and the dependence between this term and the random effects. This may lead to misleading inference and to a miss-interpretation of the baseline-related coefficients.

2.3 The joint approach

The joint distribution of $(y_{i0}, y_{i1}, \dots, y_{iT})$ is not determined from the generalized linear mixed model assumptions about the conditional model, since the marginal distribution of y_{i0} is

not specified. The joint approach involves the specification of an equation for the baseline and maximum likelihood estimates are obtained using the full set of sample observations, allowing correlation between the initial and the follow-up periods equations. Hence, the joint approach starts from the joint density of $(y_{i0}, y_{i1}, \dots, y_{iT})$

$$L_i(y_{i0}, y_{i1}, y_{i2}, \dots, y_{iT}) = \int f(y_{i0} \mid \mathbf{x}_{i0}, b_{i0}) \prod_{t=1}^T f(y_{it} \mid \mathbf{x}_{it1}, \mathbf{x}_{it2}, y_{i0}, b_{i1}, b_{i2}) h(\mathbf{b}_i) d\mathbf{b}_i \quad (7)$$

where $\mathbf{b}_i = (b_{i0}, b_{i1}, b_{i2})$ and \mathbf{x}_{i0} is a set of covariates which can be included in the linear predictor for y_{i0} .

In other words, we assume that the baseline counts are conditionally Poisson distributed random variables with canonical parameter, λ_{i0} , modeled as

$$\log(\lambda_{i0}) = \mathbf{x}'_{i0} \boldsymbol{\psi} + b_{i0} \quad (8)$$

where $\boldsymbol{\psi}$ is a vector of regression parameters and b_{i0} is an individual-specific random effect.

2.4 Computational details

To evaluate the integrals in (6) and in (7), which can be analytically solved only in few cases (see e.g. Molenberghs et al., 2010 for a review on this topic), we can use Gaussian Quadrature or Adaptive Gaussian Quadrature. Both numerical methods work well only if the random effects distribution, $h(\cdot)$, is Gaussian and as long as sufficient quadrature points are used; nevertheless, possible inaccuracy may arise for the logistic part of the model when even 20-points integration are considered (Crouch and Spiegelman, 1990). To overcome such a limitation and to relax the assumption of normally distributed random effects, an alternative is leaving $h(\cdot)$ unspecified and rely on non-parametric maximum likelihood (NPML) estimation of the mixing distribution (Laird, 1978). The NPML estimation of a mixing distribution is a discrete distribution on a finite number of mass-points, say K . We focus on the joint approach, since the naive can be viewed as as special case. Under this assumption, the $L_i(\cdot)$ function for the joint approach is given by

$$L_i(y_{i0}, y_{i1}, y_{i2}, \dots, y_{iT}) = \sum_{k=1}^K f(y_{i0} \mid \mathbf{x}_{i0}, b_{k0}) \prod_{t=1}^T f(y_{it} \mid \mathbf{x}_{it1}, \mathbf{x}_{it2}, y_{i0}, b_{k1}, b_{k2}) p_k \quad (9)$$

where the mass-points $\mathbf{b}_k = \{b_{k0}, b_{k1}, b_{k2}\}$ and the corresponding probabilities $p_k = \Pr(\mathbf{B}_i = \mathbf{b}_k) = \Pr(B_{i0} = b_{k0}, B_{i1} = b_{k1}, B_{i2} = b_{k2})$ have to be estimated together with other model parameters. Therefore, we can write down the likelihood function as

$$L(\cdot) = \prod_{i=1}^n \sum_{k=1}^K f(y_{i0} | \mathbf{x}_{i0}, b_{k0}) \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it1}, \mathbf{x}_{it2}, y_{i0}, b_{k1}, b_{k2}) p_k \prod_{i=1}^n \sum_{k=1}^K \frac{\exp(-\lambda_{i0k}) \lambda_{i0k}^{y_{i0}}}{y_{i0}!} \prod_{t=1}^T \pi_{itk}^{d_{it}} \left[(1 - \pi_{itk}) \frac{\exp(-\lambda_{itk}) \lambda_{itk}^{y_{it}}}{y_{it}! \exp(1 - \lambda_{itk})} \right]^{1-d_{it}} p_k \quad (10)$$

where $d_{it} = I(y_{it} = 0)$,

$$\text{logit}(\pi_{itk}) = \mathbf{x}'_{it1} \boldsymbol{\beta} + \alpha_1 y_{i0} + y_{i0} \mathbf{x}'_{it1} \boldsymbol{\phi}_1 + b_{k1} \quad (11)$$

$$\log(\lambda_{itk}) = \mathbf{x}'_{it2} \boldsymbol{\gamma} + \alpha_2 y_{i0} + y_{i0} \mathbf{x}'_{it2} \boldsymbol{\phi}_2 + b_{k2} \quad (12)$$

and

$$\log(\lambda_{i0k}) = \mathbf{x}'_{i0} \boldsymbol{\psi} + b_{k0}. \quad (13)$$

As can be easily noticed, the likelihood function (10) is the likelihood of a finite mixture with an unknown number of components, K , which should be estimated together with other model parameters. To obtain parameter estimates we make standard assumptions:

- K is fixed and unknown;
- y_{i0} and y_{it} are conditionally independent random variables
- there is an unobserved multivariate random variable \mathbf{B}_i following a discrete distribution with K support points with associated mass points p_k , where $p_k = \Pr(\mathbf{B}_i = \mathbf{b}_k)$ with $\sum_{k=1}^K p_k = 1$.

For fixed K we can use the following EM algorithm to obtain the maximum likelihood estimation for model parameters. We introduce an unobservable vector of component indicator variables, $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ where

$$z_{ik} = \begin{cases} 1, & \mathbf{B}_i = \mathbf{b}_k \\ 0, & \text{otherwise} \end{cases}$$

Adopting a multinomial distribution for the unobservable vector \mathbf{z}_i , we can derive the complete data log-likelihood

$$\ell_c(\cdot) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left\{ \log(p_k) + \log(f_{i0k}) + \sum_{t=1}^T \log(f_{itk}) \right\} \quad (14)$$

where

$$f_{i0k} = \frac{\exp(-\lambda_{i0k}) \lambda_{i0k}^{y_{i0}}}{y_{i0}!}$$

and

$$f_{itk} = \pi_{itk}^{d_{it}} \left[(1 - \pi_{itk}) \frac{\exp(-\lambda_{itk}) \lambda_{itk}^{y_{it}}}{y_{it}! \exp(1 - \lambda_{itk})} \right]^{1-d_{it}}.$$

Within the E-step of the algorithm, the presence of the missing data indicator in the complete data log-likelihood is handled by taking the conditional expectation of (14) given the observations and the current parameter estimates, say $\boldsymbol{\theta} = \{\boldsymbol{\delta}, \mathbf{p}\} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}_0, \mathbf{b}, \mathbf{p}\}$. In other words, at iteration r , we replace z_{ik} with its conditional expectation, say w_{ik}

$$E \left[z_{ik} \mid \mathbf{y}_i, \boldsymbol{\theta}^{(r)} \right] = w_{ik}^{(r)} = \frac{p_k^{(r)} \left[f_{i0k} \prod_{t=1}^T f_{itk} \right]}{\sum_{k=1}^K p_k^{(r)} \left[f_{i0k} \prod_{t=1}^T f_{itk} \right]}$$

where $w_{ik}^{(r)}$ is the posterior probability that unit i belongs to component k . The conditional expectation of (14) is thus given by

$$E_{\boldsymbol{\theta}^{(r)}} [\ell_c(\cdot) \mid \mathbf{y}_i] = \mathcal{Q}(\cdot) = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(r)} \left\{ \log(p_k) + \log(f_{i0k}) + \sum_{t=1}^T \log(f_{itk}) \right\}. \quad (15)$$

Conditionally on updated weights, we maximize $\mathcal{Q}(\cdot)$ with respect to $\boldsymbol{\theta}$ to obtain maximum likelihood parameter estimates

$$\frac{\partial \mathcal{Q}}{\partial p_k} = \sum_{i=1}^n \left\{ \frac{w_{ik}^{(r)}}{\hat{p}_k^{(r)}} - \frac{w_{iK}^{(r)}}{\hat{p}_K^{(r)}} \right\} = 0 \quad (16)$$

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\delta}} = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(r)} \frac{\partial}{\partial \boldsymbol{\delta}} \left[\log(f_{i0k}) + \sum_{t=1}^T \log(f_{itk}) \right] = 0 \quad (17)$$

Solving the first equation we obtain

$$\hat{p}_k^{(r)} = \sum_{i=1}^n \frac{w_{ik}^{(r)}}{n}$$

which represents a well-known result in the finite mixture context. Given $w_{ik}^{(r)}$, the score equations in (17) are weighted sums of likelihood equations for three independent generalized linear models: the baseline model, the binary model and the one for the positive counts. However, as a consequence of truncation, the sample should be augmented to consistently estimate the parameter vector for the truncated data model (details are provided in Böhning and Schön, 2005).

The E- and M- steps are repeatedly alternated until the log-likelihood difference between two iterations decreases by an arbitrary small amount.

3 Simulation study

The aim of the study is to investigate if ignoring endogeneity of the baseline variable could lead to biased estimates for the corresponding effect as well as for the effects of those covariates which are *correlated* with the baseline variable. In each simulation, 250 samples from the proposed model with $T = 3, 5, 7$ (panel length), $n = 100, 250, 500$ (sample size) and $K = 2, 3$ (number of latent classes) have been generated. Only balanced designs are considered, but the extension to unbalanced designs is straightforward. We fit the models introduced in Sections 2.2 and 2.3. Parameter estimates are obtained via maximum likelihood as described in Section 2.4 using Matlab programming. All the estimates are averaged over the number of simulations.

The samples have been generated according to the following *true* generating process

$$\log(\lambda_{i0}) = \psi_0 + \psi_1 x_{i0} + b_{k0}$$

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{it1} + \alpha_1 y_{i0} + b_{k1}$$

$$\log(\lambda_{it}) = \gamma_0 + \gamma_1 x_{it2} + \alpha_2 y_{i0} + b_{k2}$$

The covariates are independently generated from a standard Normal distribution for $i = 1, \dots, n$ and $t = 1, \dots, T$. The true values of the regression parameters are chosen as

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 0.45 \\ -0.50 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -0.30 \\ 0.25 \end{bmatrix};$$

we also define

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_0 \\ \psi_1 \end{bmatrix} = \begin{bmatrix} 0.30 \\ 0.50 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 0.50 \\ 0.25 \end{bmatrix}.$$

According to the value of K , the parameters for the latent process are chosen as follows:

$$\mathbf{K} = \mathbf{2}: \mathbf{b}_1 = (b_{10}, b_{11}, b_{12}) = (0.70; -0.70; 1.05), \mathbf{b}_2 = (b_{20}, b_{21}, b_{22}) = (-1.30; 1.30; -1.95), \\ \mathbf{p} = (p_1, p_2) = (0.65, 0.35);$$

$$\mathbf{K} = \mathbf{3}: \mathbf{b}_1 = (b_{10}, b_{11}, b_{12}) = (0.70; -0.70; 1.05), \mathbf{b}_2 = (b_{20}, b_{21}, b_{22}) = (-0.30; 0.30; -0.45), \\ \mathbf{b}_3 = (b_{30}, b_{31}, b_{32}) = (-1.30; 1.30; -1.95), \mathbf{p} = (p_1, p_2, p_3) = (0.5, 0.3, 0.2)$$

The simulation results in terms of MSE of the ML estimator of each parameter of interest (we provide only estimates of covariate coefficients and prior probabilities, \mathbf{p} , for sake of brevity) are shown in Table 1-3 (when $K = 2$) and Table 4-6 (when $K = 3$), together with the average and the interquartile range of the standard errors computed for every sample. We can observe that, with both $K = 2$ and $K = 3$, the MSE of each estimator is always moderate and decreases as n and T increase. Moreover, its standard deviation decreases at the expected rate of \sqrt{n} with respect to n and at a faster rate with respect to T . Finally, for each estimator, the average standard error is always close to the standard deviation, these standard errors also have a very low variability from sample to sample.

As can be seen, in the naive approach, the baseline estimated effects are permanently biased in the binary equation. The variance of the parameter estimates reduces with increasing sample size (n), but the bias does not, at least not substantially. Other model parameters are consistently estimated, as the adopted covariates are independent on the mechanism generating the baseline variable. However, this setting is likely unrealistic, and the behavior of the exogenous model should be better investigated to understand the effect

of ignoring endogeneity of the baseline variable when some of the adopted covariates are dependent on the baseline mechanism, as in the joint approach. In fact, taking into account the endogeneity, we find evidence indicating that the baseline estimates are more efficient than ones obtained in the naive one.

To verify the effect of model misspecification, we consider a further setting in which the individual- and outcome-specific random effects follow a multivariate continuous distribution. The simulation settings were formulated as previously discussed with the only difference that \mathbf{b}_i s are independently generated from a multivariate Gaussian random variable with mean $\boldsymbol{\mu} = [0, 0, 0]$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{02}\sigma_0\sigma_2 \\ \rho_{10}\sigma_0\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{20}\sigma_0\sigma_2 & \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Two settings are considered with respect to Σ to account for different associations among outcomes. Thus, for the correlations coefficients, the following vectors are assumed: $\rho_{jj'} = \rho_{j'j}$, $j, j' = 0, 1, 2$. For ease of discussion, we assume $\boldsymbol{\rho} = (\rho_{01}, \rho_{02}, \rho_{12}) = (-0.5, 0.5, -0.5)$ or $\boldsymbol{\rho} = (-0.25, 0.25, -0.25)$, and $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$; in this way, the random effects are dependent and the baseline variable can not be considered exogenous. Table 7 shows the simulation results concerning the MLE with $n = 100; 250$ and $T = 7$. Under this setting the number of latent classes K is unknown and we computed the MLE of parameters adopting the value of K chosen with AIC (or BIC). In most cases the number of latent classes selected with this criterion is small. Using different values for n and T , we obtain similar results, confirming that the proposed model can adequately approximate a model based on a multivariate Gaussian latent distribution. With regards to obtained results, we can observe the same consistent behavior of parameter estimates shown in Table 1-6, even if in this case, the mean square error is often greater, reflecting the greater heterogeneity present under this simulation scheme. On average, a larger sample size is necessary to obtain reliable parameter estimates, which, in the case $n = 100$, show huge variability. Table 7 shows a clear and consistent path with respect to model parameters as the sample size

increases, irrespective of the values of the correlations. This implies that the proposed model can be used even when the association between outcomes is weak or some doubts about the endogeneity of the baseline variable might arise; indirectly confirming the robustness of the proposed method.

4 An example: Outpatient psychiatric service use data

In this section we provide an application of the naive and the joint approaches using a dataset on outpatient psychiatric service use. Data are from the Access to Community Care of Effective Services and Support (ACCESS) study (Rosencheck et al., 2002). Socio-demographic characteristics, housing, income, psychiatric and substance use, as well as mental health symptoms and the use of various types of health and social services have been collected during the study.

We consider a subsample of the ACCESS study already analyzed by Neelon et al. (2010) in a Bayesian framework. This study involves 680 individuals, which self-report their utilization of any health services at baseline, 3- and 12-month follow-ups. Race, gender and an indicator of whether the subject participated in a substance abuse treatment program are recorded as well. Variables description is provided in Table 8. As the 42.72% of the observations in the follow-ups are zeros, the equidispersion assumption (i.e. mean equals variance) is not plausible and a mixed-effects hurdle model may represent an alternative to standard Poisson regression model to deal with overdispersion due to zero-inflation. Furthermore, the hurdle framework is quite established in the health economic literature, since it resembles accurately the *principal-agent* model: a decision/participation process (usually represented by a binary choice) and an utilization process.

According to the available variables (and to indications provided in Neelon et al.; 2010), we specify the following equations:

$$\begin{aligned} \text{logit}(\pi_{it}) &= \beta_0 + \beta_1 \text{Race}_{it} + \beta_2 \text{Gender}_{it} + \beta_3 \text{Treatment}_{it} \\ &+ \beta_4 \text{Visit12}_i + \beta_5 \text{Race}_i \text{Visit12}_i + \alpha_1 Y_{i0} + b_{i1} \end{aligned} \quad (18)$$

$$\begin{aligned}\log(\lambda_{it}) &= \gamma_0 + \gamma_1 Race_i + \gamma_2 Gender_i + \gamma_3 Treatment_{it} \\ &+ \gamma_4 Visit12_i + \gamma_5 Race_i Visit12_i + \alpha_2 Y_{i0} + b_{i2}\end{aligned}\tag{19}$$

and, when the joint approach is considered,

$$\log(\lambda_{i0}) = \psi_0 + \psi_1 Race_i + \psi_2 Gender_i + \psi_3 Treatment_{i0} + b_{i0}\tag{20}$$

where π_{it} , λ_{it} and λ_{i0} are defined in the Section 2.

Looking at parameter estimates, provided in Table 9, and considering a significance level of 5%, the 12-month follow-up indicator variable has a strong influence on the decision to use a health service, showing the presence of a trend effect which tends to reduce the probability of looking for a health service. The gender seems to affect the utilization process: males tend to use health services less frequently than females. We also get a quite interesting result: the treatment does not affect the utilization process, but only the decision one (posing, however, some questions on program effectiveness). Finally, the baseline mainly depends on the race: whites are more subject to health service usage than minorities (blacks/latinos). This last result represents a difference with the work of Neelon et al. (2010), which obtain a non-significant estimate of the race coefficient. Similar results arise by adopting the naive approach, but, in general, parameter estimates obtained via the joint approach are more efficient than those obtained by applying the naive approach, confirming the intuition enlightened by the simulation study.

The presence of heterogeneity between individuals leads to a number of different utilization experiences captured by the latent structure. In fact, as a by-product of the adopted estimation procedure, we identify a different number of homogeneous clusters among subjects. It is interesting to notice that in the naive approach we estimate 3 groups, while by using a joint approach we obtain 9 groups. Formally, for the random effects in the naive approach, we have the following three-points bivariate distribution $\mathbf{b} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\} = \{(2.1460; 1.0916), (-0.2465; -0.7111); (-2.3937, 0.4429)\}$ with probabilities $\boldsymbol{\pi} = \{\pi_1, \pi_2, \pi_3\} = \{0.2731; 0.5347; 0.1895\}$. In other words, we have three well defined groups: the first cluster groups subjects who have a low propensity to use service but, once they decide to use any

service they use it more than the average; the second group can be interpreted in an opposite way; while the third group can be defined as the usage group.

Similarly, in the joint approach, we have a nine-points trivariate distribution given by $\mathbf{b} = \{(-1.3004; 1.1220; 1.2443); (2.9043, -0.3020; -0.8099); (1.4923; -0.4710; -0.4331); (-2.1367; 2.9582; 0.1239); (1.8753; 1.0443; 1.1583); (-0.8312; -0.1115; -0.6548); (0.4877; -2, 8438; 0.5645); (3.1647; 0.8442; 1.2018); (2.6841; -3.2628; 0.5979)\}$ with probabilities $\boldsymbol{\pi} = \{0.0943; 0.0370; 0.1978; 0.1638; 0.0416; 0.2809; 0.1256; 0.0125; 0.0464\}$. Even if different in number the estimated groups identify behaviors close to those described by the naive approach. The main difference relies on the fact the groups are identified accounting for baseline random effects too, which strongly modified the random effects distribution, changing not only the location of the random terms but its entire distribution. As already remarked in the theoretical part of this work (see Section 2.2), the naive approach does not allow to distinguish between the impact that baseline may have on the observed follow-ups and on the random effects distribution. Therefore, the independence assumption between random effects and baseline may not be suitable. As a further confirmation of this statement, we obtain a relatively high negative correlation between the baseline and the binary process ($\rho_{b_{k0}, b_{k1}} = -0.4491$) which confirms that the need of specifying an association structure accounting, somehow, for correlations among variables. Similarly a positive correlation has been estimated between the random terms in the two parts of the hurdle model ($\rho_{b_{k1}, b_{k2}} = 0.1012$) indicating that subjects having a lower propensity toward the use of any service tend to use services more days than the average once they *decide* to use them.

5 Discussion

In this paper we discuss a mixed-effects hurdle model for zero-inflated longitudinal counts, where a baseline variable, potentially endogenous, is included in the linear predictors for the two parts describing the hurdle model. To solve the endogeneity problem, we proceed by defining a secondary model for the baseline variable, linked to the hurdle model through a

common latent structure. We propose to use a finite mixture representation for the regression model, estimating the unknown (multivariate) random effect distribution through a discrete (multivariate) distribution. By assuming that the latent process is discrete, we avoid any parametric assumptions with the advantage of permitting exact computation of the likelihood of the model without requiring quadrature or Monte Carlo methods. On the other hand, some simulation results show that the maximum likelihood estimator of the parameters has an acceptable bias even when data are generated by a continuous latent process. However, it should be stressed that the idea one has about the nature of the latent structure may be quite different from one case to another. In some cases, the probabilistic structure of the finite mixture is used only as a tool for modeling unobserved heterogeneity.

Introducing the problem of initial conditions we applied the simulation approach and showed inefficiency of the naive approach which ignores the baseline issue. Using the joint approach, the baseline is model in a generalized linear mixed model framework and consistent parameter estimates are obtained. Our overall conclusion from the performed simulation study is that modeling the baseline as an endogenous variable, we improve the performance of the model, producing more accurate estimates than the analogous based on the naive approach. The application of both approaches to real data enlightens that the baseline may influence the latent structure as well, and a different association structure between longitudinal measurements on the same individual can be specified by including in the latent structure a model which accounts explicitly for the baseline (e.g. via a logit specification of the prior probabilities).

A possible drawback in correlations estimate may arise when correlation is high, i.e. correlation tends towards the bounds of the interval $[-1;1]$. In fact, when a discrete mixing distribution is adopted, the correlation coefficient is estimated on a small number (K) of points, and, when correlation is high, this can lead to a set of almost aligned K points. Nevertheless, it should be noticed that the correlation coefficient is not a formal parameter in the proposed model, but it is rather a by-product of the adopted estimation approach.

A final point concerns possible extensions of the proposed approach. An interesting

extension consists of allowing the panel length to vary among individuals, where different mechanisms of missingness can be considered. The different types of missing-data mechanisms impact the validity of statistical analysis, and thus identification (or informed guess) of the mechanisms by which incomplete data come to be missing is crucial to select the optimal model to address data missingness, distinguishing between ignorable from non-ignorable dropout mechanisms. Another possible extensions concerns the specification of a different latent structure. The individual-specific random effects, used in the mixed-effects hurdle model to take into account unobserved heterogeneity between individuals and correlation between processes, are assumed to be time-constant. This assumption is common to many other models for longitudinal data. However, if the effect of unobservable factors is not time-constant, there can be bias in the parameter estimates and a more flexible proposal accounting for time-dependence might be considered.

References

- [1] Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251–262.
- [2] Aitkin, M. and Alfò, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modelling. *Statistical Modelling* **3**, 291–303.
- [3] Alfò, M. and Aitkin, M. (2006). Variance component models for longitudinal count data with baseline information: epilepsy data revisited. *Statistics and Computing* **16**, 231–238.
- [4] Alfò, M. and Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *The Canadian Journal of Statistics* **38**, 197–216.
- [5] Alfò, M., Maruotti, A. and Trovato, G. (2011). A finite mixture model for multivariate counts under endogenous selectivity. *Statistics and Computing* **21**, 185–202.

- [6] Auteri, M. and Maruotti, A. (2011). Modeling waiting times in the Italian National Health Service. *Applied Economics Letters*, forthcoming.
- [7] Crouch, E. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp(t^2) dt$: application to logistic-normal models. *Journal of the American Statistical Association* **85**, 464–469.
- [8] Deb, P., Munkin, M. and Trivedi, P. (2006). Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics* **21**, 1081–1099.
- [9] Davidson, R. and Mackinnon, J. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, Oxford.
- [10] Fotohui, A.R. (2005). The initial conditions problem in longitudinal binary process: A simulation study. *Simulation Modelling, Practice and Theory* **13**, 566–583.
- [11] Fotohui, A.R. (2007). The initial conditions problem in longitudinal count process: A simulation study. *Simulation Modelling, Practice and Theory* **15**, 589–604.
- [12] Greene, W. (2009). Models for count data with endogenous participation. *Empirical Economics* **36**, 133–173.
- [13] Hasan, M. T. and Sneddon, G. (2009). Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics, Simulation and Computation* **38**, 638–653.
- [14] Heckman, J.J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, in C.F. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press: Cambridge, MA, pp. 114–178.
- [15] Hsiao, C. (2003). *Analysis of Panel Data*. Second Edition. Cambridge University Press: Cambridge, UK.

- [16] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**, 805–811.
- [17] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1–19.
- [18] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.
- [19] Neelon, B.H., O’Malley, A.J. and Normand S.T. (2010). Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling* **10**, 421–439.
- [20] Rosencheck, R.A., Lam, J.A., Morrissey, J.P., Calloway, M.O., Stolar, M. Randolph, F. and the ACCESS National Evaluation Team (2002). Service Systems Integration and Outcomes for Mentally Ill Homeless Persons in the ACCESS Program. *Psychiatric Services* **53**, 958–966.
- [21] Todem, M., Zhang, Y., Ismail, A. and Sohn, W. (2010). Random effects regression models for count data with excess zeros in caries research. *Journal of Applied Statistics* **37**, 1661–1679.

n = 100 T = 3										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1	-0.022	0.530	0.281	0.504	0.017	-0.010	0.555	0.308	0.507	0.020
α_1	-0.038	0.184	0.035	0.109	0.009	-0.164	0.139	0.046	0.100	0.010
β_1	0.021	0.194	0.038	0.165	0.009	0.012	0.208	0.043	0.165	0.009
α_2	0.006	0.068	0.005	0.032	0.004	0.017	0.064	0.004	0.031	0.004
ψ_1	-0.134	0.205	0.060	0.212	0.011					
p_1	0.091	0.073	0.014			0.124	0.071	0.020		
p_2	-0.091	0.073	0.014			-0.124	0.071	0.020		

n = 100 T = 5										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1	0.030	0.397	0.158	0.387	0.012	0.032	0.402	0.162	0.389	0.011
α_1	-0.027	0.146	0.022	0.085	0.006	-0.091	0.105	0.019	0.082	0.006
β_1	0.022	0.132	0.018	0.128	0.006	0.022	0.131	0.018	0.128	0.006
α_2	0.005	0.055	0.003	0.024	0.003	0.003	0.046	0.002	0.024	0.003
ψ_1	-0.084	0.213	0.052	0.209	0.009					
p_1	0.070	0.058	0.008			0.134	0.052	0.021		
p_2	-0.069	0.058	0.008			-0.134	0.052	0.021		

n = 100 T = 7										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1	-0.004	0.337	0.113	0.324	0.008	-0.004	0.335	0.112	0.325	0.007
α_1	-0.029	0.113	0.014	0.072	0.005	-0.067	0.086	0.012	0.071	0.005
β_1	-0.003	0.118	0.014	0.107	0.006	-0.004	0.121	0.015	0.107	0.006
α_2	0.006	0.051	0.003	0.020	0.002	0.001	0.036	0.001	0.020	0.002
ψ_1	-0.092	0.186	0.043	0.211	0.007					
p_1	0.061	0.054	0.007			0.111	0.050	0.015		
p_2	-0.061	0.054	0.007			-0.111	0.050	0.015		

Table 1: Simulation results for n = 100 and K = 2

		n = 250 T = 3									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		0.026	0.319	0.103	0.317	0.009	0.034	0.330	0.110	0.320	0.009
α_1		-0.026	0.126	0.017	0.068	0.004	-0.138	0.080	0.025	0.063	0.004
β_1		0.019	0.103	0.011	0.103	0.005	0.018	0.106	0.011	0.103	0.005
α_2		0.002	0.037	0.001	0.019	0.002	0.006	0.029	0.001	0.019	0.002
ψ_1		-0.077	0.127	0.022	0.132	0.003					
p_1		0.040	0.046	0.004			0.119	0.053	0.017		
p_2		-0.040	0.046	0.004			-0.119	0.053	0.017		
		n = 250 T = 5									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		0.007	0.239	0.057	0.242	0.004	0.012	0.241	0.058	0.244	0.031
α_1		-0.019	0.093	0.009	0.053	0.003	-0.083	0.061	0.011	0.051	0.019
β_1		0.004	0.081	0.007	0.079	0.003	0.003	0.081	0.006	0.079	0.021
α_2		0.003	0.036	0.001	0.014	0.001	0.001	0.024	0.001	0.014	0.008
ψ_1		-0.095	0.127	0.025	0.132	0.004					
p_1		0.023	0.036	0.002			0.091	0.036	0.010		
p_2		-0.023	0.036	0.002			-0.091	0.036	0.010		
		n = 250 T = 7									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		0.014	0.219	0.048	0.204	0.003	0.016	0.222	0.050	0.205	0.003
α_1		-0.019	0.092	0.009	0.045	0.002	-0.052	0.047	0.005	0.045	0.002
β_1		0.000	0.068	0.005	0.067	0.002	0.000	0.066	0.004	0.067	0.002
α_2		0.005	0.038	0.001	0.012	0.001	-0.001	0.019	0.000	0.012	0.001
ψ_1		-0.101	0.118	0.024	0.132	0.003					
p_1		0.018	0.034	0.001			0.071	0.036	0.006		
p_2		-0.018	0.034	0.001			-0.070	0.036	0.006		

Table 2: Simulation results for n = 250 and K = 2

		n = 500 T = 3									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		0.009	0.248	0.062	0.224	0.003	0.017	0.254	0.065	0.226	0.004
α_1		-0.031	0.130	0.018	0.048	0.002	-0.148	0.061	0.026	0.044	0.002
β_1		0.013	0.078	0.006	0.072	0.002	0.008	0.076	0.006	0.072	0.002
α_2		0.010	0.042	0.002	0.013	0.001	0.009	0.022	0.001	0.013	0.001
ψ_1		-0.106	0.098	0.021	0.093	0.002					
p_1		0.091	0.050	0.011			0.167	0.026	0.028		
p_2		-0.091	0.050	0.011			-0.167	0.026	0.028		

		n = 500 T = 5									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		0.006	0.151	0.023	0.171	0.002	0.011	0.154	0.024	0.173	0.003
α_1		-0.008	0.058	0.003	0.038	0.001	-0.088	0.044	0.010	0.036	0.001
β_1		0.001	0.063	0.004	0.056	0.001	0.000	0.064	0.004	0.056	0.001
α_2		0.000	0.020	0.000	0.010	0.001	0.002	0.014	0.000	0.010	0.001
ψ_1		-0.098	0.092	0.018	0.093	0.002					
p_1		0.077	0.024	0.006			0.145	0.025	0.022		
p_2		-0.077	0.024	0.006			-0.145	0.025	0.022		

		n = 500 T = 7									
		Joint					Naive				
		bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
γ_1		-0.002	0.138	0.019	0.144	0.001	0.000	0.140	0.020	0.145	0.001
α_1		-0.008	0.050	0.003	0.032	0.001	-0.058	0.034	0.005	0.032	0.001
β_1		0.008	0.052	0.003	0.047	0.001	0.009	0.052	0.003	0.047	0.001
α_2		0.001	0.019	0.000	0.008	0.000	0.001	0.012	0.000	0.008	0.000
ψ_1		-0.099	0.087	0.017	0.093	0.002					
p_1		0.065	0.021	0.005			0.119	0.022	0.015		
p_2		-0.065	0.021	0.005			-0.119	0.022	0.015		

Table 3: Simulation results for n = 500 and K = 2

n = 100 T =3										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.003	0.504	0.254	0.492	0.018	0.006	0.502	0.252	0.492	0.019
α_1	-0.013	0.173	0.030	0.118	0.0123	-0.146	0.141	0.041	0.104	0.009
γ_1	0.035	0.232	0.055	0.183	0.016	0.030	0.260	0.069	0.183	0.014
α_2	0.006	0.072	0.005	0.037	0.005	0.060	0.101	0.014	0.035	0.005
ψ_1	-0.122	0.237	0.071	0.220	0.009	-0.500				
p_1	-0.044	0.084	0.009			-0.029	0.078	0.007		
p_2	0.054	0.061	0.007			0.035	0.058	0.005		
p_3	-0.009	0.069	0.005			-0.005	0.083	0.007		
n = 100 T =5										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.014	0.400	0.167	0.374	0.011	0.270	0.398	0.159	0.375	0.012
α_1	-0.019	0.099	0.010	0.090	0.005	0.395	0.099	0.021	0.084	0.006
γ_1	0.017	0.155	0.024	0.138	0.007	-0.488	0.159	0.026	0.138	0.007
α_2	-0.0081	0.036	0.001	0.027	0.003	0.271	0.064	0.004	0.026	0.004
ψ_1	-0.109	0.191	0.049	0.217	0.010					
p_1	-0.065	0.077	0.010			0.161	0.084	0.009		
p_2	0.067	0.064	0.009			0.359	0.061	0.007		
p_3	-0.002	0.054	0.003			0.480	0.068	0.005		
n = 100 T =7										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.032	0.319	0.103	0.314	0.007	0.033	0.324	0.106	0.315	0.007
α_1	-0.018	0.087	0.008	0.076	0.004	-0.072	0.080	0.012	0.073	0.004
γ_1	0.012	0.125	0.016	0.116	0.005	0.010	0.128	0.016	0.116	0.005
α_2	-0.002	0.030	0.001	0.022	0.003	0.014	0.048	0.003	0.022	0.003
ψ_1	-0.112	0.198	0.052	0.219	0.010					
p_1	-0.050	0.075	0.008			-0.034	0.079	0.007		
p_2	0.059	0.064	0.008			0.060	0.055	0.007		
p_3	-0.009	0.051	0.003			-0.025	0.060	0.004		

Table 4: Simulation results for n = 100 and K = 3

n = 250 T =3										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.005	0.322	0.104	0.306	0.007	0.010	0.329	0.108	0.307	0.007
α_1	-0.030	0.084	0.008	0.071	0.003	-0.160	0.074	0.031	0.064	0.003
γ_1	0.022	0.129	0.017	0.112	0.005	0.016	0.147	0.022	0.112	0.005
α_2	0.003	0.030	0.001	0.021	0.002	0.045	0.052	0.005	0.020	0.002
ψ_1	-0.115	0.132	0.031	0.137	0.003					
p_1	-0.042	0.080	0.008			-0.022	0.085	0.008		
p_2	0.064	0.065	0.008			0.052	0.055	0.006		
p_3	-0.022		0.003			-0.030	0.075	0.007		

n = 250 T =5										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	-0.017	0.233	0.055	0.235	0.004	-0.012	0.236	0.056	0.236	0.004
α_1	-0.031	0.058	0.004	0.055	0.002	-0.106	0.058	0.015	0.052	0.002
γ_1	0.016	0.092	0.009	0.086	0.003	0.015	0.095	0.009	0.086	0.003
α_2	0.001	0.019	0.000	0.016	0.001	0.020	0.036	0.002	0.016	0.001
ψ_1	-0.109	0.136	0.030	0.137	0.004					
p_1	-0.049	0.073	0.008			-0.025	0.081	0.007		
p_2	0.064	0.066	0.008			0.060	0.056	0.007		
p_3	-0.015	0.038	0.002			-0.035	0.057	0.004		

n = 250 T =7										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.005	0.207	0.043	0.198	0.002	0.006	0.209	0.044	0.198	0.002
α_1	-0.026	0.053	0.004	0.048	0.002	-0.078	0.050	0.009	0.046	0.002
γ_1	0.005	0.081	0.007	0.073	0.002	0.008	0.082	0.007	0.073	0.002
α_2	0.000	0.016	0.000	0.014	0.001	0.011	0.029	0.001	0.013	0.001
ψ_1	-0.117	0.123	0.029	0.137	0.003					
p_1	-0.0502	0.074	0.008			-0.026	0.073	0.006		
p_2	0.058	0.067	0.008			0.058	0.055	0.006		
p_3	-0.008	0.035	0.001			-0.032	0.049	0.003		

Table 5: Simulation results for n = 250 and K = 3

n = 500 T = 3										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	-0.008	0.232	0.054	0.216	0.003	0.002	0.235	0.055	0.217	0.003
α_1	-0.020	0.058	0.004	0.050	0.001	-0.145	0.057	0.024	0.046	0.002
γ_1	0.009	0.094	0.009	0.078	0.002	0.006	0.098	0.010	0.078	0.002
α_2	0.003	0.017	0.000	0.015	0.001	0.035	0.033	0.002	0.014	0.001
ψ_1	-0.119	0.101	0.024	0.097	0.002					
p_1	-0.048	0.080	0.009			-0.016	0.076	0.006		
p_2	0.066	0.065	0.009			0.052	0.051	0.005		
p_3	-0.018	0.048	0.003			-0.036	0.068	0.006		
n = 500 T = 5										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.021	0.1662	0.028	0.166	0.002	0.026	0.169	0.029	0.166	0.002
α_1	-0.027	0.043	0.003	0.039	0.001	-0.102	0.042	0.012	0.037	0.001
γ_1	0.001	0.069	0.005	0.060	0.001	0.002	0.070	0.005	0.060	0.001
α_2	0.003	0.014	0.000	0.011	0.001	0.019	0.028	0.001	0.011	0.001
ψ_1	-0.120	0.085	0.022	0.096	0.002					
p_1	-0.045	0.071	0.007			-0.033	0.073	0.006		
p_2	0.057	0.066	0.008			0.067	0.054	0.007		
p_3	-0.012	0.029	0.001			-0.034	0.055	0.004		
n = 500 T = 7										
	Joint					Naive				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.011	0.133	0.018	0.139	0.001	0.013	0.134	0.018	0.140	0.001
α_1	-0.029	0.034	0.002	0.033	0.001	-0.077	0.032	0.007	0.032	0.001
γ_1	-0.003	0.061	0.004	0.051	0.001	-0.001	0.062	0.004	0.051	0.001
α_2	0.002	0.011	0.000	0.009	0.001	0.009	0.012	0.000	0.009	0.000
ψ_1	-0.122	0.089	0.023	0.096	0.002					
p_1	-0.030	0.065	0.005			-0.016	0.068	0.005		
p_2	0.040	0.061	0.005			0.055	0.054	0.006		
p_3	-0.010	0.025	0.001			-0.039	0.042	0.003		

Table 6: Simulation results for n = 500 and K = 3

n = 100 T = 7										
	$\boldsymbol{\rho} = (\rho_{01}, \rho_{02}, \rho_{12}) = (-0.5, 0.5, -0.5)$					$\boldsymbol{\rho} = (\rho_{01}, \rho_{02}, \rho_{12}) = (-0.25, 0.25, -0.25)$				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	-0.022	0.313	0.098	0.302	0.006	0.003	0.312	0.097	0.302	0.006
α_1	-0.055	0.111	0.015	0.078	0.006	-0.029	0.112	0.013	0.080	0.005
γ_1	-0.003	0.217	0.047	0.148	0.006	0.008	0.220	0.048	0.149	0.006
α_2	0.041	0.103	0.012	0.033	0.005	0.020	0.115	0.014	0.035	0.005
ψ_1	-0.174	0.238	0.087	0.229	0.008	-0.178	0.240	0.089	0.229	0.007

n = 250 T = 7										
	$\boldsymbol{\rho} = (\rho_{01}, \rho_{02}, \rho_{12}) = (-0.5, 0.5, -0.5)$					$\boldsymbol{\rho} = (\rho_{01}, \rho_{02}, \rho_{12}) = (-0.25, 0.25, -0.25)$				
	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.	bias	s.e.	m.s.e.	ave.s.e.	IQR s.e.
β_1	0.003	0.190	0.036	0.190	0.003	0.006	0.196	0.038	0.191	0.003
α_1	-0.068	0.070	0.010	0.049	0.002	-0.026	0.075	0.006	0.050	0.002
γ_1	-0.005	0.117	0.014	0.092	0.002	0.003	0.144	0.021	0.093	0.003
α_2	0.045	0.087	0.010	0.020	0.002	0.021	0.081	0.007	0.021	0.001
ψ_1	-0.160	0.151	0.048	0.144	0.003	-0.168	0.142	0.048	0.144	0.004

Table 7: Simulation results for $n = 100; 250$ and $T = 7$ with multivariate Gaussian random effects

Variable	Definition	Relative frequency
Race	1 if black or latino	0.5603
Gender	1 if male	0.5985
Treatment	1 if a subject participated a treatment program	0.5015
Visit12	1 if the visit is the 12-month follow-up	

Table 8: Variables description

Variables	Naive Approach				Joint Approach					
	Binary model		Truncated model		Binary model		Truncated model		Baseline model	
	coeff	s.e.	coeff	s.e.	coeff	s.e.	coeff	s.e.	coeff	s.e.
Intercept	-0.901	0.227	1.480	0.053	-1.067	0.298	1.332	0.076	-0.477	0.175
Visit12	0.663	0.202	-0.085	0.054	0.684	0.205	-0.090	0.054		
Race	0.073	0.194	-0.029	0.049	0.137	0.196	-0.058	0.049	-0.134	0.021
RaceVisit12	-0.407	0.270	0.004	0.072	-0.427	0.272	0.003	0.072		
Gender	0.136	0.138	-0.072	0.036	0.248	0.140	-0.090	0.037	-0.083	0.178
Treatment	0.306	0.136	-0.079	0.037	0.400	0.138	-0.048	0.037	0.105	0.075
Baseline (Y_{i0})	-0.002	0.021	-0.012	0.006	-0.012	0.021	-0.013	0.006		

Table 9: Empirical application results